**Lip movements enhance speech representations and effective connectivity in
speech dorsal stream and its relationship with neurite architecture**

Lei Zhang^{1,3}, Yi Du^{1,2,3,4*}

¹ Institute of Psychology, Chinese Academy of Sciences, Beijing, China 100101

² CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai,
China 200031

³ Department of Psychology, University of Chinese Academy of Sciences, Beijing,
China 100049

⁴ Chinese Institute for Brain Research, Beijing, China 102206

* Corresponding author:

Dr. Yi Du

16 Lincui Road, Chaoyang district, Beijing, China 100101

Email: dui@psych.ac.cn

17 **Abstract**

18 Lip movements facilitate speech comprehension, especially under adverse listening
 19 conditions, but the neural mechanisms of this perceptual benefit at the phonemic and
 20 feature levels remain unclear. This fMRI study addresses this question by quantifying
 21 regional multivariate representation and network organization underlying audiovisual
 22 speech-in-noise perception. We found that valid lip movements enhanced neural
 23 representations of phoneme, place of articulation, or voicing feature of speech
 24 differentially in dorsal stream regions, including frontal speech motor areas and
 25 supramarginal gyrus. Such local changes were accompanied by strengthened dorsal
 26 stream effective connectivity. Moreover, the neurite orientation dispersion of left
 27 arcuate fasciculus, a structural basis of speech dorsal stream, predicted the visual
 28 enhancements of neural representations and effective connectivity. Our findings
 29 provide novel insight to speech science that lip movements promote both local
 30 phonemic and feature encoding and network connectivity in speech dorsal pathway
 31 and the functional enhancement is mediated by the microstructural architecture of the
 32 circuit.

33

34

35 **Introduction**

36 Speech perception becomes challenging in noisy environments and older adults (Du,
37 Buchsbaum, Grady, & Alain, 2016; L. Zhang, Fu, Luo, Xing, & Du, 2021). However,
38 in face-to-face communication, we routinely extract visual speech cues from the
39 speaker’s articulatory movements, which substantially benefits speech comprehension,
40 especially in challenging listening conditions (Ross, Saint-Amour, Leavitt, Javitt, &
41 Foxe, 2007; Sumby & Pollack, 1954) and in hearing impaired senior populations
42 (Puschmann et al., 2019). This multisensory perceptual gain has been validated across
43 speech hierarchies from isolated syllables to continuous speech, but it may operate in
44 distinct ways (Grant & Seitz, 1998). Visual speech relays correlated information about
45 “when” the speaker is saying (the timing of the acoustic signal, influencing attention
46 and perceptual sensitivity) and supplementary information about “what” the speaker is
47 saying (place and manner of articulation, constraining lexical selection) (Peelle &
48 Sommers, 2015). For continuous speech, the temporal coherence between the area of
49 mouth opening and speech envelope facilitates the attentive tracking of the speaker,
50 signals temporal markers to segment words or syllables, or provides linguistic cues,
51 thereby improving speech intelligibility (Grant & Seitz, 1998; Hauswald, Lithari,
52 Collignon, Leonardelli, & Weisz, 2018; Park, Kayser, Thut, & Gross, 2016). For
53 speech syllables and words, visual lip movements provide the place and manner of
54 articulation to constrain lexical competition (Grant & Walden, 1996). The visual
55 speech head start processed before speech vocalization is thought to increase the
56 precision of articulatory prediction (Karas et al., 2019). Growing

57 magnetoencephalography (MEG) and electroencephalogram (EEG) studies have
 58 emphasized on neural entrainment and encoding of continuous speech under the
 59 audiovisual context (Crosse, Butler, & Lalor, 2015; Crosse, Di Liberto, & Lalor, 2016;
 60 Giordano et al., 2017; Keitel, Gross, & Kayser, 2020; Park, Ince, Schyns, Thut, &
 61 Gross, 2018), that largely advances our understanding of multisensory speech
 62 processing. However, direct observation of where in the brain and how valid visual
 63 speech information modulates the focal neural representations of phonemes (the most
 64 fundamental linguistic unit) and articulatory-phonetic features, as well as the network
 65 organization during speech-in-noise perception, is still lacking. Moreover, it remains
 66 unknown which neuroanatomical structure undergirds functional changes underlying
 67 the visual enhancement of speech-in-noise perception.

68 Previous MEG and EEG research on continuous speech has shown that visual
 69 speech improves the neural tracking of speech envelope (Crosse et al., 2015;
 70 Giordano et al., 2017), and facilitates the neural encoding of both spectrotemporal and
 71 phonetic features of speech (O’Sullivan, Crosse, Di Liberto, de Cheveigné, & Lalor,
 72 2021). The visual benefit on neural tracking of speech was stronger under noisy
 73 conditions than quiet conditions, demonstrating the inverse effectiveness in
 74 audiovisual speech processing (Crosse et al., 2016). Despite the limitation of spatial
 75 resolution, recent MEG studies started to locate the brain regions involved in the
 76 visual enhancement of speech encoding, including the left motor cortex and inferior
 77 frontal gyrus (IFG) (Giordano et al., 2017; Keitel et al., 2020). The left posterior
 78 superior temporal gyrus/sulcus (pSTG/S) has been implicated as another critical

79 region in audiovisual speech integration in functional magnetic resonance imaging
80 (fMRI) studies (Erickson, Heeg, Rauschecker, & Turkeltaub, 2014; Nath &
81 Beauchamp, 2011) and intracranial EEG studies (Karas et al., 2019; Micheli et al.,
82 2020). However, the left pSTG/S is recently found to represent the common
83 redundant features of the bimodal signals, whereas left speech motor areas represent
84 the synergistic feature of them (Park et al., 2018). Moreover, the neural entrainment to
85 lip movements in the left motor cortex (Park et al., 2016) and enhanced effective
86 connectivity between frontal motor and temporal cortices (Giordano et al., 2017) were
87 correlated with the visual benefit on speech comprehension. Those findings are
88 consistent with the model suggesting that the speech dorsal stream, including the left
89 pSTG/S, supramarginal gyrus (SMG), and speech motor areas (IFG and
90 premotor/motor cortex), is involved in integrating visual and auditory speech in
91 addition to auditory and visual cortices (Bernstein & Liebenthal, 2014). Considering
92 that frontal speech motor areas are engaged to a higher extent in adverse listening
93 conditions to provide articulatory predictions to compensate for degraded bottom-up
94 speech processing (Alain, Du, Bernstein, Barten, & Banai, 2018; Du, Buchsbaum,
95 Grady, & Alain, 2014; Du et al., 2016; Nuttall, Kennedy-Higgins, Hogan, Devlin, &
96 Adank, 2016; Pickering & Garrod, 2013; Skipper, Devlin, & Lametti, 2017), we
97 hypothesized that visual lip movements would promote functional activities mainly
98 along the dorsal stream when speech is degraded by noise. Also, we are interested in
99 whether visual speech cues would shape neural representations of phonemes and
100 articulatory-phonetic features differentially in distinct regions and how the network

101 connectivity would be changed accordingly.

102 Here, we adopted the fMRI technique to specify the univariate and multivariate
 103 neural activities and effective connectivity patterns when subjects discriminated
 104 audiovisual consonant-vowel syllables under different signal-to-noise ratios (SNRs)
 105 with and without valid lip movements. Behaviorally, valid visual cues significantly
 106 improved phoneme identification via facilitating the recognition of place of
 107 articulation but not voicing regardless of SNR. Univariate analysis showed that right
 108 auditory and motor areas and bilateral visual regions were more activated when
 109 subjects were viewing valid visual cues. Multivariate pattern analysis (MVPA)
 110 revealed better neural representations of speech phonemes with valid visual cues in
 111 left speech motor areas and SMG. Interestingly, those regions exhibited distinct
 112 representational improvements by visual speech cues that the classification of voicing
 113 was enhanced in the left opercular part of IFG (IFG_{op}) while the classification of
 114 place of articulation was improved in the left inferior part of precentral gyrus (PrCG_{inf})
 115 and SMG. This is the first evidence that lip movements sharpened neural encoding of
 116 phonemes by predicting and constraining selective articulatory features in distinct
 117 dorsal stream regions. Next, we carried out the dynamic causal modeling (DCM)
 118 analysis to investigate the influence of lip movements on network organization
 119 involved in audiovisual speech perception. Bidirectional connectivity between
 120 SMG/AG (angular gyrus) and frontal speech motor areas (Broca's area and PrCG_{inf})
 121 and top-down connectivity from SMG/AG to sensory areas (auditory and visual
 122 cortices) were enhanced, while bottom-up connectivity from auditory cortex to

123 SMG/AG was inhibited with valid visual cues. These results suggest the auditory
 124 dorsal stream as a crucial pathway in audiovisual speech integration, which led us to
 125 further exam the relationship between the white matter basis of the speech dorsal
 126 stream, i.e., the arcuate fasciculus (AF) (Friederici, 2017; Hickok & Poeppel, 2007),
 127 and functional changes of visual enhancement. We used state-of-art neurite orientation
 128 dispersion and density imaging (NODDI) technique, which constructs a
 129 three-compartment tissue model with the multi-shell high angular resolution
 130 diffusion-weighted imaging (HARDI) data (Zhang, Schneider, Wheeler-Kingshott, &
 131 Alexander, 2012), to quantify the fine-grained microstructural neurite morphology of
 132 the left AF. We found that a greater visual enhancement of phoneme representations
 133 in the left IFG_{op} and a stronger visual enhancement of top-down connectivity from
 134 speech motor areas (Broca's area and PrCG_{inf}) to the auditory cortex were correlated
 135 with a higher neurite orientation dispersion of the left AF. Our findings provide novel
 136 evidence of both local phonemic and feature representations and network connectivity
 137 changes underlying the visual enhancement of speech-in-noise perception, and for the
 138 first time link individual microstructural variations of structure connectivity with
 139 functional activity during audiovisual speech processing.

140

141 **Results**

142 Participants (N = 24) were presented with 4 consonant-vowel syllables (/ba/, /da/, /pa/,
 143 /ta/) organized into 2 orthogonal articulatory features, place of articulation (bilabial:
 144 /ba/ and /pa/; lingua-dental: /da/ and /ta/) and voicing (voiced: /ba/ and /da/; voiceless:

145 /pa/ and /ta/). Syllables were embedded into a speech spectrum-shaped noise at -8, 0,
146 and 8 dB SNRs and paired with matching lip movements videos or still closed mouth
147 pictures in visual valid (VV) and visual invalid (VI) conditions, respectively (see
148 Materials and Methods). We measured whole-brain activity using fMRI while subjects
149 listened to and identified the audiovisual syllables. HARDI data were recorded in
150 addition to task fMRI.

151

152 **Lip movements improved recognition of place of articulation at the behavioral** 153 **level**

154 We replicated previous findings (Grant & Walden, 1996) that visual speech provides
155 place of articulation but not voicing to improve speech-in-noise identification. As
156 shown in Fig. 1A, the main effects of visual validity and SNR on phoneme
157 identification accuracy were both significant (visual validity: $F(1, 23) = 391.72$, $P <$
158 0.001 , $\eta_p^2 = 0.95$; SNR: $F(2, 46) = 31.43$, $P < 0.001$, $\eta_p^2 = 0.58$, repeated-measures
159 analysis of variance (ANOVA)) without a significant interaction ($F(2, 46) = 0.42$, $P =$
160 0.658 , $\eta_p^2 = 0.02$). However, valid visual cues did not promote the recognition of
161 voicing (e.g., if the stimulus /ba/ was identified as /da/, it was scored correct for
162 voicing) ($F(1, 23) = 0.00$, $P = 0.976$, $\eta_p^2 = 0.00$, Fig. 1C), which confirms that the
163 perception of voicing was determined by the auditory modality. In contrast, as shown
164 in Fig. 1B, valid visual cues significantly improved the recognition of place of
165 articulation (e.g., if the stimulus /ba/ was identified as /pa/, it was scored correct for
166 place) ($F(1, 23) = 433.83$, $P < 0.001$, $\eta_p^2 = 0.95$), and the SNR effect on the

167 identification of place was insignificant under the VV condition ($F(2, 23) = 0.13$, $P =$
 168 0.879 , $\eta_p^2 = 0.01$). This confirms that the recognition of place was determined
 169 mainly by the visual modality.

170
 171 **Fig. 1. Behavioral performance.** Mean percent of correct in identifying phonemes
 172 (A), the place of articulation feature (B) and the voicing feature (C) under visual valid
 173 (VV, blue line) and visual invalid (VI, red line) conditions.
 174

175 Furthermore, a multiple linear regression analysis took both the visual
 176 enhancement (defined as $VV - VI$) of place recognition and that of voicing
 177 recognition into account in predicting the visual enhancement of phoneme recognition.
 178 Results showed that the visual enhancement of phoneme perception was remarkably
 179 explained by the visual improvement of place recognition (Supplementary Table 1, β
 180 $= 0.89$, $P < 0.001$) but was not related with the visual benefit of voicing recognition
 181 (Supplementary Table 1, $\beta = 0.31$, $P = 0.249$).
 182

183 Lip movements enhanced brain activity in sensory and motor areas

184 Univariate analyses showed significantly increased blood oxygen level-dependent
 185 (BOLD) activity in bilateral visual areas (left middle occipital gyrus (MOG), inferior
 186 occipital gyrus (IOG), middle temporal gyrus (MTG), fusiform gyrus (FFG), inferior
 187 temporal gyrus (ITG); right IOG, MTG and MOG), right STG and right prCG in the

188 VV condition than in the VI condition (family-wise-error corrected $P(P_{fwe}) < 0.05$,
 189 Fig. 2A and Supplementary Table 2), indicating stronger engagement of visual,
 190 auditory and motor regions by valid visual speech. In contrast, brain activity in
 191 bilateral lingual gyrus and left supplementary motor area (SMA) were weaker in the
 192 VV condition than in the VI condition. Consistent with prior findings (Du et al., 2014),
 193 SNR significantly modulated activity in auditory and speech motor areas, including
 194 bilateral STG, insula, triangular part of IFG, SMA, middle cingulate cortex, left MTG,
 195 MOG, AG, right SMG and postcentral gyrus (Fig. 2B and Supplementary Table 2).
 196 No significant interaction between visual validity and SNR was found ($P_{fwe} > 0.05$).

198 **Fig. 2. Main effects of visual validity and SNR on BOLD activity.** (A) Regions
 199 where BOLD activity was modulated by visual validity (yellow: visual valid > visual
 200 invalid; blue: visual invalid > visual valid, $P_{fwe} < 0.05$). (B) Regions where BOLD
 201 activity was modulated by SNR ($P_{fwe} < 0.05$). FFG, fusiform gyrus; IFGtr, Inferior
 202 frontal gyrus, triangular part; INS, insula; IOG, inferior occipital gyrus; ITG, inferior
 203 temporal gyrus; LING, lingual gyrus; MOG, middle occipital gyrus; MTG, middle
 204 temporal gyrus; PoCG, postcentral gyrus; PrCG, precentral gyrus; SMG,
 205 supramarginal gyrus; STG, superior temporal gyrus.

206

207 **Lip movements sharpened the neural representations of speech phonemes and**
 208 **features**

209 To examine the effect of visual validity on the neural representations of speech, we

implemented MVPA in 50 individually defined anatomical regions of interest (ROIs) in both hemispheres (Fig. 3A) that were involved in audiovisual speech processing based on the previous review (Bernstein & Liebenthal, 2014). Support vector machine (SVM) classifiers were trained to decode the 4 phonemes on trial-wise fMRI response pattern ROI by ROI (see Materials and Methods).

A 2 (visual validity) $\times$ 3 (SNR) repeated-measures ANOVA found a significant improvement of classification accuracy of phonemes under the VV condition than the VI condition in the left IFG_{Op} ($F(1, 23) = 15.06$, false-discovery-rate corrected $P(P_{far}) = 0.033, \eta_p^2 = 0.40$) and the left SMG ($F(1, 23) = 13.35$, $P_{far} = 0.033, \eta_p^2 = 0.37$) (red regions in Fig. 3B and 3D), although visual validity did not influence the overall BOLD activity in those regions. Additionally, a marginally significant improvement was found in the left PrCG_{Inf} ($F(1, 23) = 10.07$, $P_{far} = 0.053, \eta_p^2 = 0.31$) and the left FFG ($F(1, 23) = 10.57$, $P_{far} = 0.053, \eta_p^2 = 0.32$) (yellow regions in Fig. 3B and 3D). The main effect of SNR and the interaction between SNR and visual validity were not significant in all ROIs ($P_{far} > 0.6$).

225

226 **Fig. 3. MVPA results.** (A) Regions of interest (ROIs) in MVPA consisted of 25 left
 227 and 25 right anatomical ROIs implicated in audiovisual speech processing. (B)
 228 Regions where phoneme classification accuracy under the visual valid (VV) condition
 229 was higher than that under the visual invalid (VI) condition (red: $P_{fdr} < 0.05$;
 230 yellow: $P_{fdr} < 0.06$). (C) Correlation between visual enhancement (VE) of MVPA
 231 classification accuracy of place in the left SMG and visual enhancement of behavioral
 232 performance for recognition of phonemes (red) and place of articulation (blue),
 233 respectively. ** $P < 0.01$, * $P < 0.05$ by Pearson's correlation. (D-F) The group mean
 234 and individual MVPA performance across SNRs in classifying phonemes (D), place
 235 of articulation (E) and voicing (F) in 4 ROIs. In panel D, * $P_{fdr} < 0.05$, † $P_{fdr} <$
 236 0.06 by repeated-measures ANOVA with FDR correction. In panel E and F, ** $P <$

0.01 , * $P < 0.05$, † $P < 0.06$, NS not significant by repeated-measures ANOVA without correction. Dash lines represent the chance level of classification. FFG, fusiform gyrus; IFG_{op}, opercular part of inferior frontal gyrus; PrCG_{inf}, inferior part of precentral gyrus; SMG, supramarginal gyrus.

Since valid visual cues improved the recognition of place but not voicing at the behavioral level, we further investigated whether this was the case at the neural level. For those 4 ROIs showing significant or marginally significant visual benefit on phoneme classification, classification accuracy was recalculated according to the voicing or place of articulation feature as the same steps used in the behavioral analysis (see Materials and Methods).

Unprecedentedly, we observed diverse patterns of visual benefit on neural representations of articulatory-phonetic features in different regions. The left IFG_{op} showed a significant visual enhancement on representing voicing ($F(1, 23) = 9.12$, $P = 0.006$, $\eta_p^2 = 0.28$) and a marginally significant visual benefit on representing place of articulation ($F(1, 23) = 4.05$, $P = 0.056$, $\eta_p^2 = 0.15$). The left FFG only had a marginally significant visual enhancement on encoding voicing ($F(1, 23) = 4.27$, $P = 0.050$). In contrast, the left SMG and left PrCG_{inf} showed a significant visual enhancement on representing place of articulation (SMG: $F(1, 23) = 4.54$, $P = 0.044$, $\eta_p^2 = 0.17$; PrCG_{inf}: $F(1, 23) = 11.51$, $P = 0.003$, $\eta_p^2 = 0.33$) but an insignificant effect on encoding voicing (SMG: $F(1, 23) = 0.67$, $P = 0.421$, $\eta_p^2 = 0.03$; PrCG_{inf}: $F(1, 23) = 2.00$, $P = 0.170$, $\eta_p^2 = 0.08$) (Fig. 3E-F).

Next, we performed the correlation analysis to investigate the relationship between neural representations and behavior performance. We found that the visual

enhancement of place classification in the left SMG was positively correlated with behavioral visual enhancement of place recognition (Pearson's $r = 0.61$, $P = 0.002$), so as for phoneme identification (Pearson's $r = 0.44$, $P = 0.033$) (Fig. 3C). No other correlation was found for any region (all Pearson's $|r| < 0.32$, $P > 0.125$).

Lip movements tightened the connection between dorsal stream areas and sensory cortices

We further conducted the DCM analysis to explore the effect of visual validity on the effective connectivity among audiovisual speech processing areas (Fig. 4A). Based on the univariate and MVPA results as well as the previous review (Bernstein & Liebenthal, 2014), the left SMG/AG and left speech motor areas (IFG_{op} and PrCG_{inf}) were selected as amodal hub regions in the dorsal stream, and the left auditory cortex and visual cortex were included as sensory areas (see Materials and Methods).

As shown in Fig. 4B and Supplementary Table 3, valid visual cues increased bidirectional connectivity between SMG/AG and speech motor areas (SMG/AG to speech motor areas: $t(23) = 3.06$, $P_{fdr} = 0.015$; speech motor areas to SMG/AG: $t(23) = 2.43$, $P_{fdr} = 0.041$) and top-down modulation from SMG/AG to auditory cortex ($t(23) = 3.18$, $P_{fdr} = 0.013$) and visual cortex ($t(23) = 6.06$, $P_{fdr} < 0.001$). However, bottom-up connectivity from auditory cortex to SMG/AG was inhibited by valid visual cues ($t(23) = -3.46$, $P_{fdr} = 0.011$). In addition, self-inhibition increased in auditory cortex ($t(23) = 2.54$, $P_{fdr} = 0.037$) and speech motor areas ($t(23) = 3.91$, $P_{fdr} = 0.006$), but decreased in visual cortex ($t(23) = -3.37$, $P_{fdr} = 0.011$) and

283 SMG/AG ($t(23) = -2.62$, $P_{fdr} = 0.035$) when visual cues became valid. These results
 284 indicate that auditory cortex and speech motor areas became less sensitive to inputs
 285 from other regions while visual cortex and SMG/AG became more sensitive to inputs
 286 from other regions with valid visual cues. Correlation analysis found that
 287 self-inhibition in auditory cortex was negatively correlated with behavioral visual
 288 enhancement of phonemes (Pearson's $r = -0.54$, $P = 0.006$) and behavioral visual
 289 enhancement of place recognition (Pearson's $r = -0.53$, $P = 0.008$); self-inhibition in
 290 visual cortex was negatively correlated with behavioral visual enhancement of place
 291 recognition (Pearson's $r = -0.46$, $P = 0.025$); and connectivity from speech motor
 292 areas to SMG/AG was negatively correlated with behavioral visual enhancement of
 293 voicing recognition (Pearson's $r = -0.41$, $P = 0.044$). No other connectivity-behavior
 294 correlation was found (all Pearson's $|r| < 0.4$, $P > 0.052$).

295 **Fig. 4. Dynamic causal modelling results.** (A) Regions of interest in DCM. (B) The
 296 effective connectivities that were modulated by visual validity. Solid lines: $p_{fdr} <$
 297 0.05, dashed lines: $p_{fdr} > 0.05$. Numbers represent averaged parameter estimates. AC,
 298 auditory cortex; AG, angular gyrus; IFG_{op}, opercular part of inferior frontal gyrus;
 299 PrCG_{inf}, inferior part of precentral gyrus; SMG, supramarginal gyrus; VC, visual
 300 cortex.
 301

302

303 **Left AF microstructure predicted functional visual benefits**

304 To further explore the structure-function relationship, we dissected 3 segments (long,
305 anterior and posterior segments) of the left AF from the HARDI data. The mean
306 fractional anisotropy (FA) from the diffusion tensor imaging (DTI) model and the
307 mean neurite density index (NDI) and orientation dispersion index (ODI) from the
308 NODDI model were calculated for each fiber bundle. The 3 AF segments connect
309 ROIs in the MVPA results and DCM (see Materials and Methods), making it
310 reasonable to implement correlation analyses between functional results and
311 corresponding structural indexes.

312 As shown in Fig. 5, we found that the ODI of the long segment of AF (lAF,
313 directly connecting IFG_{op} / PrCG_{inf} and posterior STG/MTG) was positively
314 correlated with the visual enhancement of MVPA phoneme classification accuracy in
315 the left IFG_{op} (Pearson's $r = 0.52$, $P = 0.010$) and the visual enhancement of
316 connectivity from speech motor areas (IFG_{op}/PrCG_{inf}) to the auditory cortex
317 (Pearson's $r = 0.41$, $P = 0.044$). Meanwhile, the FA of the left lAF showed a negative
318 correlation with the visual enhancement of MVPA phoneme classification accuracy in
319 the left IFG_{op} (Pearson's $r = -0.43$, $P = 0.032$) but no relationship with the visual
320 enhancement of connectivity from speech motor areas to the auditory cortex
321 (Pearson's $r = -0.31$, $P = 0.14$). This result was consistent with the relationship
322 between FA and ODI that FA decreases with a larger orientation variability.
323 Additionally, the visual enhancement of phoneme classification accuracy in the left
324 IFG_{op} was positively correlated with the visual enhancement of connectivity from

speech motor areas to the auditory cortex (Pearson's $r = 0.45$, $P = 0.026$). No other correlation between functional indices and structural indices of AF segments (all Pearson's $|r| < 0.23$, $P > 0.278$), nor between behavioral performance and structural indices (all Pearson's $|r| < 0.29$, $P > 0.164$) was found.

Fig. 5. Correlations between the left AF microstructural properties and functional activity. The visual enhancement of phoneme classification accuracy in the left IFG_{op} correlated with the orientation dispersion index (ODI) and fractional anisotropy (FA) of the left long segment of arcuate fasciculus (IAF) and the visual enhancement of effective connectivity from speech motor areas to the auditory cortex. The visual enhancement of effective connectivity from speech motor areas to auditory cortex correlated with the ODI of the left IAF. Solid lines: $P < 0.05$, dashed lines: $P > 0.05$, ** $P < 0.01$, * $P < 0.05$ by Pearson's correlation.

Discussion

Visual lip movements improve speech-in-noise perception likely by constraining lexical interpretations and increasing the precision of prediction of the timing and content of the upcoming speech signal (Peelle & Sommers, 2015). Here we found that

visual lip movements aided the behavioral recognition of the place of articulation rather than the voicing feature to promote phoneme perception in noisy conditions. MVPA results indicated that valid visual cues sharpened phoneme representations in speech motor areas (including the left IFG_{op} and PrCG_{inf}) and the left SMG. Although it was not significant in the behavioral results, visual information enhanced the specificity of neural representations of voicing in the left IFG_{op}, and improved the specificity of place representations in the left PrCG_{inf} and SMG. This fills the gap of knowledge that distinct dorsal stream regions are involved in processing different articulatory-phonetic features of audiovisual speech. In addition to regional neural representations, DCM analysis revealed significant visual modulations on the effective connectivity of the audiovisual speech network. With valid visual cues, SMG showed stronger bidirectional connectivity with speech motor areas and greater feedback connectivity to auditory and visual areas. Last but not least, the visual enhancement of phoneme specificity in the left IFG_{op} and the visual enhancement of effective connectivity from speech motor areas to the auditory cortex were correlated with each other, and both were correlated with the neurite architecture (orientation dispersion index) of the white matter tract connecting speech motor areas with auditory regions (the long segment of left AF). Our findings provide the first evidence that visual lip movements sharpen the neural representations of phonemes according to the place of articulation or voicing feature distinctively and promote the directed connectivity in the dorsal stream of speech processing, which is involved in sensorimotor integration (Hickok & Poeppel, 2007), and such visual benefits are

365 mediated by neurite architecture of the dorsal stream fiber tract.

366 During speech processing, the speech motor areas (including Broca's area and the
367 left ventral premotor cortex) and the left pSTG/S are two candidate regions implicated
368 in audiovisual integration (Pelle & Sommers, 2015). However, in our results, these
369 two brain regions in the right hemisphere, rather than in the left hemisphere, showed
370 stronger brain activation with valid visual cues. The stronger activation does not equal
371 to greater speech encoding ability, as only left speech motor areas showed greater
372 specificity of speech representations with valid visual cues. This is consistent with
373 previous findings that although both areas in the left could carry temporal information
374 from auditory and visual modalities (Micheli et al., 2020), the left pSTG/S only
375 represents redundant information of audiovisual speech, while the left motor areas
376 represent synergistic information of audiovisual speech (Park et al., 2018). Similarly,
377 MEG studies have found that both lip movements and speech envelope are tracked
378 better only in left speech motor areas, but not in the left pSTG/S, with valid visual
379 speech (Giordano et al., 2017; Park et al., 2016). Furthermore, a neuroimaging
380 meta-analysis showed that the left pSTG/S is more steadily activated during
381 conflicting audiovisual speech processing rather than validating speech processing
382 (Erickson et al., 2014). Therefore, the left pSTG/S may be involved in solving the
383 conflict between information from auditory and visual modalities, which was absent
384 in the current study. In contrast, left speech motor areas may underlie the improved
385 speech-in-noise perception with visual lip movements by enhancing neural
386 representations of speech.

387 Consistent with previous findings (Grant & Walden, 1996), we found that lip
388 movements significantly improved the identification of the place of articulation
389 feature but not the voicing feature of speech. This supports the notion that the
390 recognition of place and voicing is determined by the visual and auditory modality,
391 respectively. Although we can easily discriminate voiced and voiceless consonants
392 presented in noise merely by ear (90-100% correct, Fig. 1C), visual speech cues
393 enhanced the neural encoding of voicing in Broca's area (IFG_{op}) without remarkable
394 behavioral benefit. In contrast, the neural representations of place got improved by
395 visual speech cues significantly in the left ventral premotor cortex (PrCG_{inf}) and the
396 left SMG, and marginally significant in the left IFG_{op}. Although an audiovisual
397 integration effect on encoding 19-dimensional phonetic features has been observed in
398 a recent EEG study using continuous speech (O'Sullivan et al., 2021), to our
399 knowledge, this is the first study that found a distinct visual enhancement effect of
400 articulatory-phonetic feature representations in different brain regions. In the
401 dual-stream model of speech perception, the dorsal stream can be further divided into
402 the dorsal-dorsal stream that terminates in the premotor cortex (BA6, 8), and the
403 dorsal-ventral stream that terminates in Broca's area (IFG_{op}, BA44) (Friederici, 2017;
404 Rauschecker, 2018), implying the functional disassociation of the two speech motor
405 areas. Speech production studies have found that the ventral premotor cortex
406 represents articulatory gestures to a greater extent than phonemes, while Broca's area
407 represents both articulatory gestures and phonemes (Lotte et al., 2015; Mugler et al.,
408 2018). It is posit that Broca's area formulates the articulatory code which is passed to

409 the premotor and motor cortices that subsequently implement the articulation during
 410 speech production (Basilakos, Smith, Fillmore, Fridriksson, & Fedorenko, 2018;
 411 Flinker et al., 2015; Long et al., 2016). In parallel, speech motor areas are
 412 hypothesized to generate articulatory predictions to compensate for degraded speech
 413 representations in the auditory cortex during speech perception when listening context
 414 (e.g., noisy, distorted speech) requires (Alain et al., 2018; Du et al., 2014, 2016;
 415 Nuttall et al., 2016; Pickering & Garrod, 2013; Skipper et al., 2017). Moreover, the
 416 left ventral premotor cortex exhibits articulator-specific engagement in speech
 417 perception (Liang & Du, 2018; Schomers & Pulvermüller, 2016). Under the
 418 audiovisual speech perception context, visual speech constrains the lexical
 419 competition by providing articulatory gestures, especially when the visual speech
 420 head start is processed before acoustic vocalization in most cases so that the
 421 articulatory prediction could be more precise (Karas et al., 2019; Peelle & Sommers,
 422 2015). In particular, the left ventral premotor cortex is implicated in encoding the
 423 bottom-up lip movements (Ozker, Yoshor, & Beauchamp, 2018) besides receiving the
 424 top-down motor plans from Broca's area, and extracting the synergistic feature of
 425 multimodal information (Park et al., 2018). Combining these findings, we
 426 hypothesized that Broca's area and the left ventral premotor cortex might play a
 427 different role in audiovisual speech-in-noise perception. Specifically, Broca's area
 428 might launch covert rehearsal and articulatory prediction to a greater extent and
 429 higher precision with visual speech cues, which would improve the neural
 430 differentiation of both voicing representations and place representations in Broca's

431 area. On the other hand, place of articulation rather than voicing is the major
432 articulatory feature that visual lip movements provide to promote speech processing
433 (Grant & Walden, 1996). Therefore, the left premotor cortex is assumed to
434 automatically decipher the place of articulation information in lip movements by
435 recruiting premotor subregions that control corresponding articulators, leading to
436 enhanced topographical representations of place of articulation along the premotor
437 strip. However, very few studies except Callen and his colleagues (Callan, Jones, &
438 Callan, 2014) have investigated the distinct BOLD activity of the subregions in
439 speech motor areas during audiovisual speech perception, further studies are needed
440 to explore the detailed roles of speech motor subregions.

441 SMG is a multimodal brain region that anatomically and functionally connects
442 auditory, visual and speech motor areas (Bernstein & Liebenthal, 2014; Binkofski,
443 Klann, & Caspers, 2016; Donaldson, Rinehart, & Enticott, 2015). According to the
444 dual-stream model of speech perception, the left SMG maps sensory representations
445 into articulatory representations in speech processing (Gow, 2012; Hickok & Poeppel,
446 2007). Besides the speech processing network, SMG is also involved in the visual
447 dorsal stream, which processes visuomotor sequences such as eye movements and
448 hand movements (Basilakos et al., 2018; Meister, Wilson, Deblieck, Wu, & Iacoboni,
449 2007; Rauschecker, 2018). When speech is presented with lip movements, the left
450 SMG would integrate visuomotor and acoustic information to promote the
451 sensory-to-motor mapping. Therefore, the left SMG was recognized as a hub region in
452 audiovisual speech perception where neural representations of phonemes and place of

453 articulation were improved by lip movements. Notably, the visual enhancement of
454 place representations in the left SMG predicted behavioral visual enhancement of
455 phoneme recognition performance, which indicates that the representational changes
456 in the left SMG may serve as a key neural substrate of the audiovisual benefit in
457 speech processing.

458 The adding of visual speech cues not only sharpened speech representations in
459 dorsal stream regions, but also strengthened the bidirectional effective connectivity of
460 the dorsal speech stream (between SMG/AG and speech motor areas) and the
461 top-down modulation from multimodal SMG/AG to unimodal sensory areas (auditory
462 and visual cortices). The visual speech-induced stronger effective connectivity of the
463 dorsal stream implies a greater extent of sensorimotor integration, which provides
464 articulatory predictions to constrain phonological representations in sensory areas, to
465 promote speech-in-noise perception (Du et al., 2014, 2016; Du & Zatorre, 2017;
466 Hickok, Houde, & Rong, 2011; Hickok & Poeppel, 2007; Pickering & Garrod, 2013).
467 An MEG study indeed found that the behavioral visual benefit is not predicted by
468 changes in local speech entrainment but rather by enhanced effective connectivity
469 between inferior frontal and temporal cortices (Giordano et al., 2017). In the current
470 study, although we found no significant correlation between behavioral benefit and
471 frontal-temporal connectivity, we revealed a positive correlation between the neural
472 specificity of phonemes in the left IFG_{op} and the effective connectivity from speech
473 motor areas to the auditory cortex. This result suggests that the better speech
474 representations in frontal motor areas may lead to a stronger top-down constraint to

475 auditory speech processing.

476 Another finding from the DCM analysis is that when a valid visual speech cue
477 was presented, the auditory area became more self-inhibited (less sensitive to inputs
478 from the network), but the visual area became less self-inhibited (more sensitive to
479 inputs from other brain regions). This echoes a previous study (Nath & Beauchamp,
480 2011), in which the functional connectivity between the sensory cortex and the
481 multisensory area is found reliability-weighted, that the multisensory region tends to
482 be more strongly connected to the sensory area with more reliable information. That is,
483 with valid visual cues, auditory information became less dominant to speech
484 processing, and the visual cortex became more engaged in speech perception. Besides
485 self-inhibition, the feedforward connectivity from the auditory area to the multimodal
486 SMG/AG also reduced under the visual valid condition, supporting that auditory
487 inputs became less weighted when visual cues are informative. Although we did not
488 find a significant visual modulation effect regarding the connectivity from the visual
489 cortex to other brain regions, weaker self-inhibition of the visual cortex correlated
490 with stronger behavioral visual enhancement of place recognition, again
491 demonstrating the increased contribution of visual modality.

492 Importantly, we further used in vivo NODDI technique to quantify the
493 microcircuitry in terms of axon and dendrite complexity of the left AF, which is
494 recognized as the neuroanatomic foundation of the dorsal stream in speech processing
495 (Friederici, 2017; Hickok & Poeppel, 2007) and speech-in-noise perception (Li,
496 Zatorre, & Du, 2021; Tremblay et al., 2019). The structure-function correlation

analysis showed that the visual enhancement of effective connectivity from speech motor areas to the auditory cortex and the visual enhancement of phoneme representations in the left IFG_{op} were positively predicted by the ODI of the long segment of left AF, which directly connects the auditory cortex and speech motor areas. We also found a negative correlation between DTI-derived FA of the long segment of left AF and phoneme representations in the left IFG_{op}. The opposite pattern between ODI and FA is consistent with our knowledge that the larger FA is correlated with greater NDI and lower ODI (Zhang et al., 2012). Note that, NODDI has been widely used in clinical populations, and previous studies have revealed that NODDI-derived ODI and NDI of white matter (Fu et al., 2020) and grey matter (Nazeri et al., 2015; Vogt et al., 2020) provide more specific microstructural indices than DTI-derived FA and macrostructural changes to cognitive aging, mild cognitive impairment and Alzheimer's disease. The more robust structure-function correlation observed by ODI than by FA in the current study supports the above notion. However, NODDI has very recently been introduced to human cognitive neuroscience to investigate the relationship between brain morphometry and cognition in normal participants. One study has found a correlation between higher neurite density of the left planum temporale and higher temporal precision and shorter latency of auditory speech perception (Ocklenburg et al., 2018). In other two studies using HARDI data to estimate the fiber orientation distributions (FOD), the apparent fiber density (AFD) and the number of fiber orientations (NuFO) of the left AF are correlated with speech-in-noise perception criterion (Tremblay et al., 2019), and the AFD of the right

519 AF is associated with EEG effective connectivity along the AF (Oestreich, Randeniya,
520 & Garrido, 2019). To the best of our knowledge, this is the first study to introduce
521 white matter neurite imaging to speech processing research and to investigate the
522 relationship among neural representations, effective connectivity, and fiber neurite
523 architecture during audiovisual speech perception. Although our analyses were rather
524 exploratory, our findings imply that the higher dendritic complexity of the left AF
525 may contribute to stronger benefits from the visual speech in enhancing neural
526 specificity of phoneme representations and effective connectivity of the speech dorsal
527 stream. This is the first evidence of the microstructural underpinning of functional
528 performance in audiovisual speech-in-noise perception, and opens new avenue for
529 future research.

530 Lastly, we did not find a significant interaction between visual validity and SNR
531 on either BOLD activity or MVPA classification accuracy, which is unexpected since
532 the visual benefit is assumed to be stronger in more noisy conditions than quieter
533 conditions, i.e., the inverse effectiveness in audiovisual speech processing (Crosse et
534 al., 2016). This may be caused partly by stringent correction procedure for multiple
535 comparisons, and inappropriate SNR range to display the inverse effectiveness, as the
536 performance even at the highest SNR (8 dB) in the visual invalid (auditory only)
537 condition was relatively poor (~ 60% correct) inside the scanner.

538 In summary, we demonstrate that the speech dorsal stream is the key in visual
539 enhancement of speech perception in noisy environments. Lip movements enhance
540 both the specificity of phoneme representations and network connectivity of the

dorsal stream to improve speech-in-noise perception. At the feature level, the visual enhancement on encoding place of articulation is revealed in the left ventral premotor cortex and multisensory SMG, while the visual enhancement on encoding voicing is observed in Broca's area, providing novel evidence on interpreting finer roles of dorsal stream regions in articulatory-to-acoustic mapping during audiovisual speech processing. Importantly, this is the first report that the neurite orientation dispersion along the left AF can predict the visual benefits of neural representations and connectivity in the speech dorsal stream, pinpointing the microstructural property undergirding functional dynamics in multisensory speech processing. Our study paves the way for exploring local neural representations at different speech hierarchies, network dynamics, and microstructural characteristics underlying audiovisual speech perception.

553

554 **Materials and Methods**

555 **Participants**

Twenty-four young adults (19-28 years old, 12 females) participated in this study. All participants were healthy, right-handed, native Chinese speakers with no history of neurological disorder and normal hearing (average pure-tone threshold < 20 dB HL for 250 to 8,000 Hz) at both ears. All participants had signed the written consent approved by the Institute of Psychology, Chinese Academy of Sciences.

561

562 **Experimental design**

563 The stimuli comprised 4 naturally pronounced consonant-vowel syllables (/ba/, /da/,
564 /pa/, /ta/) uttered by a young Chinese female. The 4 syllables have 2 orthogonal
565 articulatory features, voicing (voiced: /ba/ and /da/; unvoiced: /pa/ and /ta/) and place
566 of articulation (bilabial: /ba/ and /pa/; lingua-dental: /da/ and /ta/). The utterances were
567 videotaped by a Sony FDR-AX45 camera in a soundproof room. Then, they were
568 digitized and edited on the computer to produce a 1-second video. Video digitizing
569 was done at 29.97 frames/s in 1024 × 768 pixels. The pictures of the videos were cut,
570 retaining the mouth and the neck part. The audio syllable stimuli were nearly 400ms
571 in duration, low-pass filtered (4-kHz), and matched for average root-mean-square
572 sound pressure level (SPL). The masker was a speech spectrum-shaped noise (4-kHz
573 low-pass, 10-ms rise-decay envelope) that was representative of the spectrum of 113
574 different sentences by 50 Chinese young female speakers. The speech stimuli were
575 presented at 90 dB SPL, and the SPL of the maskers was adjusted to produce different
576 SNRs (-8, 0, and 8 dB). Audio stimuli were presented via MRI-compatible
577 Sensimetrics S14 insert earphones (Sensimetrics Corporation) with Comply foam tips,
578 which maximally attenuate scanner noise by 40 dB.

579 The experiment was a 3 (SNR: -8, 0 and 8 dB) × 2 (visual validity: valid and
580 invalid) factor design. Matching lip movements videos and still lip pictures (the first
581 frames of the matching videos) were presented with speech signals in the VV and VI
582 conditions, respectively. In the fMRI scanner, subjects were instructed to listen to the
583 speech signals, watch the mouth on the screen, and identify the syllables by pressing
584 the corresponding button using their right-hand fingers (index to little fingers in

585 response to /ba/, /da/, /pa/, and /ta/ in half of the subjects or the reverse order in the
586 other half). Each subject completed 4 blocks of VI conditions and 4 blocks of VV
587 conditions. The conditions were arranged in an ABBA or BAAB order, which was
588 counterbalanced across participants. Each block contained 60 stimuli (20 trials $\times$ 3
589 SNRs), which were pseudo-randomly presented with an average inter-stimuli-interval
590 of 5 s (4–6 s, 0.5 s step). Stimuli were presented via Psychtoolbox (Brainard, 1997).

591

592 **Behavioral analysis**

593 We performed repeated-measures ANOVAs to investigate the effects of SNR and
594 visual validity on phoneme-syllable identification or articulatory feature identification
595 (voicing and place of articulation). Greenhouse–Geisser correction would be
596 performed if the sphericity assumption was violated. Consistent with the previous
597 study(Grant & Walden, 1996), if syllable /ba/ was recognized as /pa/, the response
598 was correct for place and incorrect for voicing, while if syllable /ba/ was recognized
599 as /da/, the response was correct for voicing, and incorrect for place. We further used
600 a multiple regression analysis to determine the contributions of the visual
601 enhancement on voicing and the visual enhancement on place to the visual
602 enhancement on phoneme recognition. Visual enhancement was defined as the
603 difference between the accuracy under the VV condition and the VI condition.
604 Statistical analysis was conducted in R (R Core Team, 2017) with the package bruceR
605 (Bao, 2020) and visualized using the package ggplot2 (Wickham, 2009).

606

607 **Functional imaging data acquisition and preprocessing**

608 Functional MRI data were collected by a 3T MRI system (Siemens Magnetom Trio)
 609 with a 20-channel head coil. T1 weighted images were acquired using the
 610 magnetization-prepared rapid acquisition gradient echo (MPRAGE) sequence (TR =
 611 2200 ms, TE = 3.49 ms, FOV = 256 mm, voxel size = 1×1×1 mm). T2 weighted
 612 images were acquired using the multiband-accelerated echo planar imaging (EPI)
 613 sequence (acceleration factor = 4, TR = 640 ms, TE = 30 ms, slices = 40, FOV = 192,
 614 voxel sizes = 3×3×3 mm).

615 The fMRI data were preprocessed using Analysis of Functional NeuroImages
 616 (AFNI) software (Cox, 1996). The first 8 volumes were removed for each block. For
 617 univariate analysis, the following preprocessing steps included slice timing, motion
 618 correction, aligning the functional image with anatomy, spatial normalization
 619 (MNI152 space), spatial smoothing with 6 mm FWHM isotropic Gaussian kernel, and
 620 scaling each voxel time series to have a mean of 100. The fMRI data were not
 621 spatially normalized, smoothed, and scaled for MVPA at the preprocessing steps.

622

623 **Univariate analysis**

624 We conducted single-subject multiple-regression modeling using the AFNI program
 625 3dDeconvolve. Six conditions of 4 syllables and 6 regressors corresponding to motion
 626 parameters were entered into the analysis. TRs were censored if the motion
 627 derivatives exceeded 0.3. For each SNR and visual validity, the four syllables were
 628 grouped and contrasted against the baseline.

629 We performed the group level analysis using the AFNI program 3dMVM. Two
630 within-subject factors (visual validity, SNR) and their interaction were put into the
631 model. Multiple comparisons were corrected using 3dClustSim (“fixed” version) with
632 real smoothness of data estimated by 3dFWHMx (acf method) (Cox, Chen, Glen,
633 Reynolds, & Taylor, 2017). 10000 Monte Carlo simulations were performed to get the
634 cluster threshold ($\alpha = 0.05$ FWE corrected, uncorrected voxel-wise $P < 0.005$).
635 Results were visualized onto an inflated cortical surface using SUMA with AFNI.

636

637 **ROI-based MVPA**

638 We implemented MVPA in anatomically defined ROIs specific to each participant,
639 thus no spatial normalization and smoothing was applied. We chose anatomical
640 ROI-based MVPA rather than searchlight MVPA because we wished to preserve
641 borders between spatially adjacent areas (e.g., IFG and STG) that were found to
642 exhibit differential phoneme specificity at noisy conditions (Du et al., 2014, 2016).
643 Freesurfer’s automatic anatomical parcellation (aparc2009 (Destrieux, Fischl, Dale, &
644 Halgren, 2010)) algorithm was used to define a set of 148 cortical and subcortical
645 ROIs from the individual’s anatomical image. We further divided STG into equational
646 anterior and posterior portions, and divided prCG into equational dorsal and ventral
647 parts. 25 ROIs in the left hemisphere that were closely related to audiovisual speech
648 perception (Bernstein & Liebenthal, 2014) and the 25 counterparts in the right
649 hemisphere were intersected with the Freesurfer mask to generate the 50 ROIs for
650 MVPA. The classifiers were trained using SVM algorithm with a linear kernel. The

cost parameter C was set to 1. The input feature was univariate trial-wise β coefficients that were estimated using AFNI program 3dLSS, which was recommended performing MVPA in fast event-related designs (Mumford, Turner, Ashby, & Poldrack, 2012). For each condition, the first level analysis of ROI-based MVPA was conducted within each anatomical ROI using the Decoding Toolbox (Hebart, Görden, & Haynes, 2015). Twenty-fold cross-validation was used to evaluate classification performance, which was measured by the mean accuracy. Each fold contained a β coefficient of 1 trial of each syllable. We then conducted repeated-measures ANOVA with within-subject factors of visual validity and SNR in each ROI. Multiple comparisons were corrected with an FDR $q = 0.05$ using Benjamini–Hochberg procedure.

To further investigate the potentially different feature encoding visual benefits in different ROIs, for ROIs that showed significant or marginally significant visual enhancement on phoneme classification after FDR correction, we recalculated the classification accuracy according to the voicing and place feature with the same approach as the behavioral analysis. Repeated-measures ANOVA with within-subject factors of visual validity and SNR was performed to examine which feature representation was visually enhanced in each ROI.

DCM analysis

We used DCM (Friston, Harrison, & Penny, 2003) analysis in SPM12 to assess effective connectivity among brain regions involved in audiovisual speech processing.

673 Based on prior knowledge from the literature (Bernstein & Liebenthal, 2014) and our
 674 univariate and MVPA results, 4 ROIs (speech motor areas including IFG_{op} and
 675 PrCG_{inf}, SMG and AG, auditory cortex and visual cortex) in the left hemisphere that
 676 were critical in audiovisual speech processing were selected in the DCM analysis.
 677 Although IFG_{op} and PrCG_{inf} showed different visual enhancement of feature
 678 representations in MVPA results, IFG_{op} and PrCG_{inf} were combined into one
 679 speech motor ROI in order to simplify the DCM model complexity. We identified the
 680 coordinates of each ROI according to the peak voxel of that region in the group-level
 681 activation under the VV condition. The group mean coordinates of ROIs were
 682 IFG_{op}/PrCG_{inf} (-60, 6, 24), SMG/AG (-54, -52, 42), auditory cortex (-52, -18, 8) and
 683 visual cortex (-24, -94, -6).

684 We extracted the time series of each ROI according to the guideline
 685 (https://en.wikibooks.org/wiki/SPM/Timeseries_extraction). Since variation showing
 686 the maximum effect of interest between participants existed, we defined individual
 687 ROI as an 8 mm sphere centered on the individual peak activation voxel within a 15
 688 mm sphere centered on the group peak voxel. This approach allowed individual ROIs
 689 to have slight variation between subjects, and be close to group peak coordinates.
 690 Voxels within the individual ROIs survived with the $p < 0.05$ uncorrected threshold
 691 were used to exclude the noisiest voxels within the ROIs. As suggested by the
 692 developers (Zeidman et al., 2019), if subjects with no voxel survived in an ROI
 693 existed, we increased the threshold with the step of 0.05 until all subjects got survived
 694 voxels in the ROI. Finally, we extracted the time series of survived voxels and used

695 the first principal component of the extracted time series within the ROI in the
696 subsequent DCM analysis.

697 We specified the modeling according to the DCM guide (Zeidman et al., 2019).
698 DCM models the change of a neuronal signal x using the following bilinear state
699 equation:

$$\dot{x} = Ax + \sum_{j=1}^m u_j B^j x + Cu$$

700 Matrix A denotes endogenous connectivity between modeled regions during baseline.
701 Matrix B^j denotes the rate of change (in Hz) in connectivity between modelled
702 regions with the j -th modulatory inputs. Matrix C represents how neuronal activity
703 was influenced by the stimulus inputs. In the current study, the interested matrix was
704 B^j that represents how effective connectivity among audiovisual speech processing
705 regions was changed with valid visual cues in contrast to invalid cues. A positive
706 parameter indicates that the connectivity increased. Conversely, a negative parameter
707 indicates that the connectivity decreases. In addition, diagonal elements of the matrix,
708 which indicate the intrinsic within-region self-inhibition, were also switched on. The
709 more positive the self-connection parameter, the more inhibited the region, so the less
710 it will respond to inputs from the network.

711 Since we assumed that valid visual speech would alter almost all the connections,
712 we estimated a fully connected DCM for each subject using Bayesian model inversion.
713 The main effect of the task (all trials) was set as the driving input to all ROIs (matrix
714 C). The main effect of visual validity and SNR were set as modulatory inputs on the
715 self-inhibition of each ROI (diagonal elements of matrix B^j) and between-ROI

connections (non-diagonal elements of matrix B^j). Then, the `spm_dcm_peb` function was used to update the individual subject' parameters using the group-level connection strengths as empirical priors, making summary statistics optimal. Finally, we extracted the expected connectivity parameters of matrix B^j from all participants. A one-sample t-test and FDR correction were performed to investigate the statistical significance of modulation ($P_{fdr} < 0.05$).

Diffusion-weighted imaging data acquisition and preprocessing

Diffusion-weighted imaging (DWI) data were collected with following parameters: TR = 4000 ms, TE = 79 ms, voxel size = $1.5 \times 1.5 \times 1.5$ mm, FOV = 192 mm, 64 gradient directions with two b values of 1000 s/mm^2 and 2000 s/mm^2 , and 5 acquisitions without diffusion weighting ($b = 0 \text{ s/mm}^2$), which yielded the HARDI data.

The DWI data were pre-processed using MRtrix3 and FSL software (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012; Tournier et al., 2019). Preprocessing steps included denoising, unringing, eddy current and motion correction, and bias field correction using the N4 algorithm provided in Advanced Normalization Tools. Gradient directions were also corrected after eddy current and motion correction.

HARDI tractography

Following the preprocessed step, tractography was conducted by MRtrix3 (Tournier et al., 2019). Firstly, 3-tissue (white matter, grey matter, and cerebrospinal fluid)

738 response functions were obtained by the command “dwi2response dhollander”
 739 (Dhollander, Mito, Raffelt, & Connelly, 2019). Secondly, based on the response
 740 functions and preprocessed DWI data, we carried out multi-shell multi-tissue
 741 constrained spherical deconvolution (CSD) to estimate the FOD of each voxel.
 742 Thirdly, we performed the whole-brain tractography using second-order integration
 743 over FOD (IFOD2) probabilistic algorithm. Ten million streamlines were generated
 744 for each subject. Lastly, the command “tcksift” was used to filter the 10 million
 745 streamlines to 1 million streamlines (Smith, Tournier, Calamante, & Connelly, 2013).

746 AF has three segments: the long segment corresponds to the classical AF directly
 747 connecting the Broca's area and the Wernicke's area, the indirect anterior segment
 748 connects the Broca's area and the Geschwind's territory (the inferior parietal lobule)
 749 and the indirect posterior segment connects the Geschwind's territory and the
 750 Wernicke's area (Catani, Jones, & Ffytche, 2005). Three ROIs in the left brain
 751 (Broca's area: IFG_{op} and PrCG_{inf}; the Geschwind's territory: SMG and AG;
 752 Wernicke's area: posterior STG and MTG) were extracted from individual anatomical
 753 image parceled by Freesurfer, which were also used in MVPA. The extracted 3 ROIs
 754 were used to dissect the 3 segments of left AF according to the definition above in the
 755 native space.

756

757 **NODDI and DTI indexes calculation and correlation analysis**

758 A NODDI model was fitted to each voxel of the preprocessed DWI data using
 759 AMICO python toolbox (Daducci et al., 2015) for each subject. The NODDI model is

760 based on a 3-compartment tissue model (intra-cellular, extra-cellular and
761 cerebrospinal fluid) and provides 3 indexes that are more specific to the white matter
762 microstructure properties than FA index from the tensor model. NDI describes the
763 number of neurites within a voxel, and ODI represents the variability of neurite
764 orientations. For comparison, we also fitted a DTI model to each voxel of the
765 preprocessed data using MRtrix3, generating an FA map for each subject. The
766 relationship between FA and NODDI indexes is that FA increases with the increase of
767 NDI or the decrease of ODI, and vice versa (Zhang et al., 2012).

768 We extracted the mean FA, NDI, and ODI along each AF segment for the correlation
769 analysis. We performed the Shapiro-Wilk normality test to determine whether the
770 variable was normally distributed. Then, we calculated Pearson's correlation
771 coefficients to assess the relationship between functional and structural results
772 because all variables were normally distributed.

773

774 **Acknowledgments**

775 **Funding:** This research was supported by grants from the National Natural Science
776 Foundation of China (Grant No. 31671172 and 31822024), and the Strategic Priority
777 Research Program of Chinese Academy of Sciences (Grant No. XDB32010300) to Y.
778 Du.

779 **Author contributions:** L. Zhang acquired and analyzed the data. L. Zhang and Y. Du
780 designed the experiment, interpreted the results and wrote the manuscript.

781 **Competing interests:** The authors declare no competing financial interests.

783 **References**

- 784 Alain, C., Du, Y., Bernstein, L. J., Barten, T., & Banai, K. (2018). Listening under
785 difficult conditions: An activation likelihood estimation meta-analysis. *Human*
786 *Brain Mapping*, 39(7), 2695–2709. Retrieved from
787 <https://doi.org/10.1002/hbm.24031>
- 788 Bao, H.-W.-S. (2020). bruceR: BRoadly Useful Collections and Extensions of R
789 functions. Retrieved from <https://github.com/psychbruce/bruceR>
- 790 Basilakos, A., Smith, K. G., Fillmore, P., Fridriksson, J., & Fedorenko, E. (2018).
791 Functional Characterization of the Human Speech Articulation Network.
792 *Cerebral Cortex*, 28(5), 1816–1830. Retrieved from
793 <https://doi.org/10.1093/cercor/bhx100>
- 794 Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech
795 perception. *Frontiers in Neuroscience*, 8, 1–18. Retrieved from
796 <https://doi.org/10.3389/fnins.2014.00386>
- 797 Binkofski, F. C., Klann, J., & Caspers, S. (2016). Chapter 4 - On the Neuroanatomy
798 and Functional Role of the Inferior Parietal Lobule and Intraparietal Sulcus. In G.
799 Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 35–47). San Diego:
800 Academic Press. Retrieved from
801 <https://doi.org/https://doi.org/10.1016/B978-0-12-407794-2.00004-3>
- 802 Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10(4), 433–436.
803 Retrieved from <https://doi.org/10.1163/156856897X00357>
- 804 Callan, D. E., Jones, J. A., & Callan, A. (2014). Multisensory and modality specific

805 processing of visual speech in different regions of the premotor cortex. *Frontiers*
806 *in Psychology*, 5, 389. Retrieved from <https://doi.org/10.3389/fpsyg.2014.00389>

807 Catani, M., Jones, D. K., & Ffytche, D. H. (2005). Perisylvian language networks of
808 the human brain. *Annals of Neurology*, 57(1), 8–16. Retrieved from
809 <https://doi.org/10.1002/ana.20319>

810 Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional
811 magnetic resonance neuroimages. *Computers and Biomedical Research*, 29(3),
812 162–173. Retrieved from <https://doi.org/10.1006/cbmr.1996.0014>

813 Cox, R. W., Chen, G., Glen, D. R., Reynolds, R. C., & Taylor, P. A. (2017). FMRI
814 Clustering in AFNI: False-Positive Rates Redux. *Brain Connectivity*, 7(3),
815 152–171. Retrieved from <https://doi.org/10.1089/brain.2016.0475>

816 Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances
817 cortical entrainment to continuous auditory speech in noise-free conditions.
818 *Journal of Neuroscience*, 35(42), 14195–14204. Retrieved from
819 <https://doi.org/10.1523/JNEUROSCI.1829-15.2015>

820 Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2016). Eye can hear clearly now:
821 Inverse effectiveness in natural audiovisual speech processing relies on
822 long-term crossmodal temporal integration. *Journal of Neuroscience*, 36(38),
823 9888–9895. Retrieved from <https://doi.org/10.1523/JNEUROSCI.1396-16.2016>

824 Daducci, A., Canales-Rodríguez, E. J., Zhang, H., Dyrby, T. B., Alexander, D. C., &
825 Thiran, J. P. (2015). Accelerated Microstructure Imaging via Convex
826 Optimization (AMICO) from diffusion MRI data. *NeuroImage*, 105, 32–44.

827 Retrieved from <https://doi.org/10.1016/j.neuroimage.2014.10.026>

828 Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of
829 human cortical gyri and sulci using standard anatomical nomenclature.
830 *NeuroImage*, 53(1), 1–15. Retrieved from
831 <https://doi.org/10.1016/j.neuroimage.2010.06.010>

832 Dhollander, T., Mito, R., Raffelt, D., & Connelly, A. (2019). Improved white matter
833 response function estimation for 3-tissue constrained spherical deconvolution.
834 *Proc. Intl. Soc. Mag. Reson. Med*, (May 11-16), 555. Retrieved from
835 [https://www.researchgate.net/publication/331165168_Improved_white_matter_r](https://www.researchgate.net/publication/331165168_Improved_white_matter_response_function_estimation_for_3-tissue_constrained_spherical_deconvolution)
836 [esponse_function_estimation_for_3-tissue_constrained_spherical_deconvolution](https://www.researchgate.net/publication/331165168_Improved_white_matter_response_function_estimation_for_3-tissue_constrained_spherical_deconvolution)

837 Donaldson, P. H., Rinehart, N. J., & Enticott, P. G. (2015). Noninvasive stimulation
838 of the temporoparietal junction: A systematic review. *Neuroscience and*
839 *Biobehavioral Reviews*, 55, 547–572. Retrieved from
840 <https://doi.org/10.1016/j.neubiorev.2015.05.017>

841 Du, Y., Buchsbaum, B. R., Grady, C. L., & Alain, C. (2014). Noise differentially
842 impacts phoneme representations in the auditory and speech motor systems.
843 *Proceedings of the National Academy of Sciences of the United States of*
844 *America*, 111(19), 7126–31. Retrieved from
845 <https://doi.org/10.1073/pnas.1318738111>

846 Du, Y., Buchsbaum, B. R., Grady, C. L., & Alain, C. (2016). Increased activity in
847 frontal motor cortex compensates impaired speech perception in older adults.
848 *Nature Communications*, 7, 1–12. Retrieved from

849 <https://doi.org/10.1038/ncomms12241>

850 Du, Y., & Zatorre, R. J. (2017). Musical training sharpens and bonds ears and tongue
851 to hear speech better. *Proceedings of the National Academy of Sciences of the*
852 *United States of America*, 114(51), 13579–13584. Retrieved from
853 <https://doi.org/10.1073/pnas.1712223114>

854 Erickson, L. C., Heeg, E., Rauschecker, J. P., & Turkeltaub, P. E. (2014). An ALE
855 meta-analysis on the audiovisual integration of speech signals. *Human Brain*
856 *Mapping*, 35(11), 5587–5605. Retrieved from <https://doi.org/10.1002/hbm.22572>

857 Flinker, A., Korzeniewska, A., Shestyuk, A. Y., Franaszczuk, P. J., Dronkers, N. F.,
858 Knight, R. T., & Crone, N. E. (2015). Redefining the role of broca’s area in
859 speech. *Proceedings of the National Academy of Sciences of the United States of*
860 *America*, 112(9), 2871–2875. Retrieved from
861 <https://doi.org/10.1073/pnas.1414491112>

862 Friederici, A. D. (2017). Evolution of the neural language network. *Psychonomic*
863 *Bulletin and Review*, 24(1), 41–47. Retrieved from
864 <https://doi.org/10.3758/s13423-016-1090-x>

865 Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling.
866 *NeuroImage*, 19(4), 1273–1302. Retrieved from
867 [https://doi.org/10.1016/S1053-8119\(03\)00202-7](https://doi.org/10.1016/S1053-8119(03)00202-7)

868 Fu, X., Shrestha, S., Sun, M., Wu, Q., Luo, Y., Zhang, X., ... Ni, H. (2020).
869 Microstructural White Matter Alterations in Mild Cognitive Impairment and
870 Alzheimer’s Disease: Study Based on Neurite Orientation Dispersion and

871 Density Imaging (NODDI). *Clinical Neuroradiology*, 30(3), 569–579. Retrieved
872 from <https://doi.org/10.1007/s00062-019-00805-0>

873 Giordano, B. L., Ince, R. A. A., Gross, J., Schyns, P. G., Panzeri, S., & Kayser, C.
874 (2017). Contributions of local speech encoding and functional connectivity to
875 audio-visual speech perception. *eLife*, 6, e24763. Retrieved from
876 <https://doi.org/10.7554/eLife.24763>

877 Gow, D. W. (2012). The cortical organization of lexical knowledge: A dual lexicon
878 model of spoken language processing. *Brain and Language*, 121(3), 273–288.
879 Retrieved from <https://doi.org/10.1016/j.bandl.2012.03.005>

880 Grant, K. W., & Seitz, P. F. (1998). Measures of auditory–visual integration in
881 nonsense syllables and sentences. *The Journal of the Acoustical Society of*
882 *America*, 104(4), 2438–2450.

883 Grant, K. W., & Walden, B. E. (1996). Evaluating the articulation index for
884 auditory–visual consonant recognition. *The Journal of the Acoustical Society of*
885 *America*, 100(4), 2415–2424. Retrieved from <https://doi.org/10.1121/1.417950>

886 Hauswald, A., Lithari, C., Collignon, O., Leonardelli, E., & Weisz, N. (2018). A
887 visual cortical network for deriving phonological information from intelligible
888 lip movements. *Current Biology*, 28(9), 1453–1459.

889 Hebart, M. N., Görden, K., & Haynes, J. D. (2015). The decoding toolbox (TDT): A
890 versatile software package for multivariate analyses of functional imaging data.
891 *Frontiers in Neuroinformatics*, 8, 88. Retrieved from
892 <https://doi.org/10.3389/fninf.2014.00088>

893 Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor Integration in Speech
894 Processing: Computational Basis and Neural Organization. *Neuron*, 69(3),
895 407–422. Retrieved from <https://doi.org/10.1016/j.neuron.2011.01.019>

896 Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing.
897 *Nature Reviews Neuroscience*, 8(5), 393–402. Retrieved from
898 <https://doi.org/10.1038/nrn2113>

899 Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M.
900 (2012). FSL. *NeuroImage*, 62, 782–790.

901 Karas, P. J., Magnotti, J. F., Metzger, B. A., Zhu, L. L., Smith, K. B., Yoshor, D., &
902 Beauchamp, M. S. (2019). The visual speech head start improves perception and
903 reduces superior temporal cortex responses to auditory speech. *eLife*, 8, e48116.
904 Retrieved from <https://doi.org/10.7554/eLife.48116>

905 Keitel, A., Gross, J., & Kayser, C. (2020). Shared and modality-specific brain regions
906 that mediate auditory and visual word comprehension. *eLife*, 9, 1–23. Retrieved
907 from <https://doi.org/10.7554/ELIFE.56972>

908 Li, X., Zatorre, R. J., & Du, Y. (2021). The Microstructural Plasticity of the Arcuate
909 Fasciculus Undergirds Improved Speech in Noise Perception in Musicians.
910 *Cerebral Cortex*, 31(9), 3975–3985. Retrieved from
911 <https://doi.org/10.1093/cercor/bhab063>

912 Liang, B., & Du, Y. (2018). The functional neuroanatomy of lexical tone perception:
913 An activation likelihood estimation meta-analysis. *Frontiers in Neuroscience*, 12,
914 495. Retrieved from <https://doi.org/10.3389/fnins.2018.00495>

915 Long, M. A., Katlowitz, K. A., Svirsky, M. A., Clary, R. C., Byun, T. M. A., Majaj,
916 N., ... Greenlee, J. D. W. (2016). Functional Segregation of Cortical Regions
917 Underlying Speech Timing and Articulation. *Neuron*, 89(6), 1187–1193.
918 Retrieved from <https://doi.org/10.1016/j.neuron.2016.01.032>

919 Lotte, F., Brumberg, J. S., Brunner, P., Gunduz, A., Ritaccio, A. L., Guan, C., &
920 Schalk, G. (2015). Electrocorticographic representations of segmental features in
921 continuous speech. *Frontiers in Human Neuroscience*, 9, 97. Retrieved from
922 <https://doi.org/10.3389/fnhum.2015.00097>

923 Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The
924 Essential Role of Premotor Cortex in Speech Perception. *Current Biology*,
925 17(19), 1692–1696. Retrieved from <https://doi.org/10.1016/j.cub.2007.08.064>

926 Micheli, C., Schepers, I. M., Ozker, M., Yoshor, D., Beauchamp, M. S., & Rieger, J.
927 W. (2020). Electrocorticography reveals continuous auditory and visual speech
928 tracking in temporal and occipital cortex. *European Journal of Neuroscience*,
929 51(5), 1364–1376. Retrieved from <https://doi.org/10.1111/ejn.13992>

930 Mugler, E. M., Tate, M. C., Livescu, K., Templer, J. W., Goldrick, M. A., & Slutzky,
931 M. W. (2018). Differential representation of articulatory gestures and phonemes
932 in precentral and inferior frontal gyri. *Journal of Neuroscience*, 38(46),
933 9803–9813. Retrieved from <https://doi.org/10.1523/JNEUROSCI.1206-18.2018>

934 Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving
935 BOLD activation in event-related designs for multivoxel pattern classification
936 analyses. *NeuroImage*, 59(3), 2636–2643. Retrieved from

937 <https://doi.org/10.1016/j.neuroimage.2011.08.076>

938 Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal
939 sulcus connectivity during perception of noisy audiovisual speech. *Journal of*
940 *Neuroscience*, 31(5), 1704–1714. Retrieved from
941 <https://doi.org/10.1523/JNEUROSCI.4853-10.2011>

942 Nazeri, X., Chakravart, M., Rotenberg, D. J., Rajji, T. K., Rathi, X., Michailovich, O.
943 V., & Voineskos, A. N. (2015). Functional consequences of neurite orientation
944 dispersion and density in humans across the adult lifespan. *Journal of*
945 *Neuroscience*, 35(4), 1753–1762. Retrieved from
946 <https://doi.org/10.1523/JNEUROSCI.3979-14.2015>

947 Nuttall, H. E., Kennedy-Higgins, D., Hogan, J., Devlin, J. T., & Adank, P. (2016).
948 The effect of speech distortion on the excitability of articulatory motor cortex.
949 *NeuroImage*, 128, 218–226. Retrieved from
950 <https://doi.org/10.1016/j.neuroimage.2015.12.038>

951 O’Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., de Cheveigné, A., & Lalor, E. C.
952 (2021). Neurophysiological indices of audiovisual speech processing reveal a
953 hierarchy of multisensory integration effects. *Journal of Neuroscience*, 41(23),
954 4991–5003. Retrieved from <https://doi.org/10.1523/JNEUROSCI.0906-20.2021>

955 Ocklenburg, S., Friedrich, P., Fraenz, C., Schlüter, C., Beste, C., Güntürkün, O., &
956 Genç, E. (2018). Neurite architecture of the planum temporale predicts
957 neurophysiological processing of auditory speech. *Science Advances*, 4(7),
958 eaar6830. Retrieved from <https://doi.org/10.1126/sciadv.aar6830>

- 959 Oestreich, L. K. L., Randeniya, R., & Garrido, M. I. (2019). Auditory white matter
960 pathways are associated with effective connectivity of auditory prediction errors
961 within a fronto-temporal network. *NeuroImage*, 195, 454–462. Retrieved from
962 <https://doi.org/10.1016/j.neuroimage.2019.04.008>
- 963 Ozker, M., Yoshor, D., & Beauchamp, M. S. (2018). Frontal cortex selects
964 representations of the talker’s mouth to aid in speech perception. *eLife*, 7, e30387.
965 Retrieved from <https://doi.org/10.7554/eLife.30387>
- 966 Park, H., Ince, R. A. A., Schyns, P. G., Thut, G., & Gross, J. (2018). Representational
967 interactions during audiovisual speech entrainment: Redundancy in left posterior
968 superior temporal gyrus and synergy in left motor cortex. *PLoS Biology*, 16(8),
969 e2006558. Retrieved from <https://doi.org/10.1371/journal.pbio.2006558>
- 970 Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the
971 observers’ low-frequency brain oscillations to facilitate speech intelligibility.
972 *eLife*, 5, e14521. Retrieved from <https://doi.org/10.7554/eLife.14521>
- 973 Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual
974 speech perception. *Cortex*, 68, 169–181. Retrieved from
975 <https://doi.org/10.1016/j.cortex.2015.03.006>
- 976 Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production
977 and comprehension. *Behavioral and Brain Sciences*, 36(4), 329–347. Retrieved
978 from <https://doi.org/10.1017/S0140525X12001495>
- 979 Puschmann, S., Daeglau, M., Stropahl, M., Mirkovic, B., Rosemann, S., Thiel, C. M.,
980 & Debener, S. (2019). Hearing-impaired listeners show increased audiovisual

981 benefit when listening to speech in noise. *Neuroimage*, 196, 261–268.

982 R Core Team. (2017). R: A Language and Environment for Statistical Computing.

983 Vienna, Austria. Retrieved from <https://www.r-project.org/>

984 Rauschecker, J. P. (2018). Where, When, and How: Are they all sensorimotor?

985 Towards a unified view of the dorsal pathway in vision and audition. *Cortex*, 98,

986 262–268. Retrieved from <https://doi.org/10.1016/j.cortex.2017.10.020>

987 Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do

988 you see what I am saying? Exploring visual enhancement of speech

989 comprehension in noisy environments. *Cerebral Cortex*, 17(5), 1147–1153.

990 Retrieved from <https://doi.org/10.1093/cercor/bhl024>

991 Schomers, M. R., & Pulvermüller, F. (2016). Is the sensorimotor cortex relevant for

992 speech perception and understanding? An integrative review. *Frontiers in*

993 *Human Neuroscience*, 10, 435. Retrieved from

994 <https://doi.org/10.3389/fnhum.2016.00435>

995 Skipper, J. I., Devlin, J. T., & Lametti, D. R. (2017). The hearing ear is always found

996 close to the speaking tongue: Review of the role of the motor system in speech

997 perception. *Brain and Language*, 164, 77–105. Retrieved from

998 <https://doi.org/10.1016/j.bandl.2016.10.004>

999 Smith, R. E., Tournier, J. D., Calamante, F., & Connelly, A. (2013). SIFT:

1000 Spherical-deconvolution informed filtering of tractograms. *NeuroImage*, 67,

1001 298–312. Retrieved from <https://doi.org/10.1016/j.neuroimage.2012.11.049>

1002 Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in

1003 Noise. *Journal of the Acoustical Society of America*, 26(2), 212–215. Retrieved
1004 from <https://doi.org/10.1121/1.1907309>

1005 Tournier, J. D., Smith, R., Raffelt, D., Tabbara, R., Dhollander, T., Pietsch, M., ...
1006 Connelly, A. (2019). MRtrix3: A fast, flexible and open software framework for
1007 medical image processing and visualisation. *NeuroImage*, 202, 116–137.
1008 Retrieved from <https://doi.org/10.1016/j.neuroimage.2019.116137>

1009 Tremblay, P., Perron, M., Deschamps, I., Kennedy-Higgins, D., Houde, J. C., Dick, A.
1010 S., & Descoteaux, M. (2019). The role of the arcuate and middle longitudinal
1011 fasciculi in speech perception in noise in adulthood. *Human Brain Mapping*,
1012 40(1), 226–241. Retrieved from <https://doi.org/10.1002/hbm.24367>

1013 Vogt, N. M., Hunt, J. F., Adluru, N., Dean, D. C., Johnson, S. C., Asthana, S., ...
1014 Bendlin, B. B. (2020). Cortical Microstructural Alterations in Mild Cognitive
1015 Impairment and Alzheimer’s Disease Dementia. *Cerebral Cortex*, 30(5),
1016 2948–2960. Retrieved from <https://doi.org/10.1093/cercor/bhz286>

1017 Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag
1018 New York. Retrieved from <http://ggplot2.org>

1019 Zeidman, P., Jafarian, A., Corbin, N., Seghier, M. L., Razi, A., Price, C. J., & Friston,
1020 K. J. (2019). A guide to group effective connectivity analysis, part 1: First level
1021 analysis with DCM for fMRI. *NeuroImage*, 200, 174–190. Retrieved from
1022 <https://doi.org/10.1016/j.neuroimage.2019.06.031>

1023 Zhang, H., Schneider, T., Wheeler-Kingshott, C. A., & Alexander, D. C. (2012).
1024 NODDI: Practical in vivo neurite orientation dispersion and density imaging of

1025 the human brain. *NeuroImage*, 61(4), 1000–1016. Retrieved from
 1026 <https://doi.org/10.1016/j.neuroimage.2012.03.072>
 1027 Zhang, L., Fu, X., Luo, D., Xing, L., & Du, Y. (2021). Musical Experience Offsets
 1028 Age-Related Decline in Understanding Speech-in-Noise: Type of Training Does
 1029 Not Matter, Working Memory Is the Key. *Ear and Hearing*, 42(2), 258–270.
 1030 Retrieved from <https://doi.org/10.1097/AUD.0000000000000921>
 1031
 1032

Supplementary Materials

Supplementary Table 1. Results of the multiple linear regression analysis using the visual enhancement of recognition accuracy according to the place of articulation and voicing feature to predict the visual enhancement of phoneme identification accuracy.

	Visual enhancement (phoneme)
(Intercept)	0.02
Visual enhancement (place)	0.89***
Visual enhancement (voicing)	0.31
R^2	0.87
Adj. R^2	0.85
Num. obs.	24

* $p < .05$, ** $p < .01$, *** $p < .001$

Supplementary Table 2. Brain regions showing significant difference between visual valid and visual invalid conditions and a significant main effect of SNR ($P_{fwe} < 0.05$). AG, angular gyrus; Bi, bilateral; CALG, calcarine gyrus; CUN, cuneus; FFG, fusiform gyrus; HG, Heschl gyrus; IFGtr, Inferior frontal gyrus, triangular part; INS, insula; IOG, inferior occipital gyrus; ITG, inferior temporal gyrus; L, left; LING, lingual gyrus; MCC, middle cingulate cortex; MFG, middle frontal gyrus; MOG, middle occipital gyrus; MTG, middle temporal gyrus; PoCG, postcentral gyrus; PrCG, precentral gyrus; R, right; SFGmed, superior frontal gyrus, medial; SMA, supplementary motor area; SMG, supramarginal gyrus; STG, superior temporal gyrus.

Brain Regions	Peak MNI coordinates			Peak t/F-value	No. of voxel
	x	y	z		
VV>VI					
R IOG/MTG/MOG/FFG/ITG	30	-90	-9	9.68	1046
L MOG/IOG/MTG	-33	-99	-6	10.1	944
L FFG/ITG	-42	-45	-21	6.64	132
R PrCG/MFG	54	3	57	4.77	59
R STG	78	-45	18	4.66	40
VV<VI					
Bi LING/CALG/CUN	-9	-54	-3	-6.55	1399
Bi SMA/L SFGmed	0	21	54	-4.21	58
SNR main effect					
R STG/HG	48	-21	6	27.48	435
L STG	-60	-36	12	19.30	239
L INS/IFGtr	-33	21	0	19.78	139
R INS/IFGtr	33	27	9	17.84	132
L MTG/MOG/AG	-45	-69	18	13.02	124
L Caudate Nucleus	-6	6	15	33.85	114
Bi SMA/SFGmed/MCC	6	33	33	9.30	113
R Putamen	27	-3	-9	16.88	87
L Putamen	-30	-3	0	12.32	85
R SMG/PoCG	60	-21	42	10.19	42

1055 **Supplementary Table 3. DCM results. Parameter estimates of the modulation**
 1056 **effect of visual validity on the effective connectivity, the t and P values of one**
 1057 **sample t-tests, and P values after FDR correction. AC, auditory cortex; AG,**
 1058 **angular gyrus; IFGop, opercular part of inferior frontal gyrus; PrCGinf, inferior**
 1059 **part of precentral gyrus; SMG, supramarginal gyrus; VC, visual cortex.**
 1060

	Estimate (mean)	Estimate (SEM)	t	P	P_{fdr}
From AC to					
AC	0.149	0.059	2.54	0.019	0.037
VC	0.053	0.088	0.60	0.553	0.590
SMG/AG	-0.198	0.057	-3.46	0.002	0.011
IFGop/PrCGinf	-0.107	0.056	-1.92	0.068	0.109
From VC to					
AC	0.142	0.087	1.63	0.117	0.170
VC	-0.136	0.040	-3.37	0.003	0.011
SMG/AG	0.006	0.065	0.09	0.926	0.926
IFGop/PrCGinf	0.176	0.132	1.33	0.196	0.241
From SMG/AG to					
AC	0.187	0.059	3.18	0.004	0.013
VC	0.325	0.054	6.06	<0.001	<0.001
SMG/AG	-0.172	0.066	-2.62	0.016	0.035
IFGop/PrCGinf	0.312	0.102	3.06	0.006	0.015
From IFGop/PrCGinf to					
AC	-0.046	0.071	-0.65	0.521	0.590
VC	-0.087	0.062	-1.41	0.171	0.228
SMG/AG	0.165	0.068	2.43	0.023	0.041
IFGop/PrCGinf	0.184	0.047	3.91	0.001	0.006

A **B** **C**

A

Main effect of visual validity

B

Main effect of SNR

A

B